Assignment – Based Questions:

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   - Spring has the lowest demand. Fall has the highest demand followed by summer.
   - Year 2019 had a significant increase in the demand compared to 2018.
   - Less demand on holidays.
   - Working day or not, the demand seems to be the same.
   - Demand increased during summers and fall starting from Jun till October. September being the highest. The demand is less during the year end and start.
   - Weekday doesn't give a clear picture.
   - The demand is clearly high on days with clear weather conditions. Extremely low on days with light rain and thunderstorms. There are no records for days with Heavy-Rain conditions.
2. **Why is it important to use drop_first=True during dummy variable creation?**
   For a category, if there are n levels then we need only n-1 dummy variables. As the condition with all zeros would act as the base state and would imply that if nothing else is true then that particular condition is true. So creating another variable for that base condition becomes redundant.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   'Temp' has the highest correlation coefficient with target variable 'cnt' of 0.63.
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - Calculate the residuals given by : y(actual) – y(predicted)
   - Plot a histogram for residuals.
   - The plot should show a normal distribution.
   - The plot was centred at 0.0
   - Plot a scatter plot for error terms.
   - No visible pattern observed in the scatter plot. The error terms were randomly scattered.
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - Temp with coeff 0.43
   - Yr with coeff 0.23
   - Weathersit_Ligh-Rain with coeff -0.29

General – Subjective Questions

1. **. Explain the linear regression algorithm in detail.**
   Linear Regression is used when we want to find out the effect of some predictor variables on a target variable provided there is a linear relationship between them.

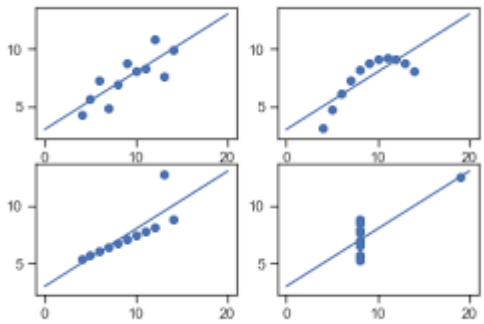
Since there is a linear relationship therefore we try to fit the best possible straight line between the target and the predictor variables and their relationship is explained by the equation:

$$y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots\ldots B_nX_n$$

Steps followed in building a linear regression module:

- Reading, Understanding & Visualising the Data
- Pre-Processing the Data (handling categorical variables, Train-Test Split, Scalling)
- Training the model using train data.
- Residual Analysis
- Predict & Evaluate the model on test data.

2. **Explain the Anscombe's quartet in detail.**
   Anscombe's quartet consists of 4 data sets that have identical simple descriptive statistics like mean, standard deviation, correlation coeffecients, R-sqaured values, linear regression line but when plotted graphically they are completely different data sets.



3. **What is Pearson's R?**
   It is a measure of correlation between two variables. Also known as correlation coefficient. It can hold values between -1 to 1. A negative number indicating that if one variable increases the other decreases i.e. they are inversely related. A positive number on the other hand indicates a direct relation between the two, meaning if one variable increases the other also increases.
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Scaling is a technique used for bringing all the variables to a comparable scale. If one variable is in 10s, the other in hundreds, another one in thousands and if we do not rescale then some coefficients would be either very large or very small compared to other which could hamper the evaluation of the model.
   There are 2 ways in which we can rescale:
   1. Min-Max Scaling also known as Standardised Scaling
   2. Normalised Scaling

   Difference between both:

- Normalised centres the mean at 0.0 with a standard deviation of 1 whereas Standardised compresses the values between 0 and 1.
- In Normalised scaling the values are not bounded whereas in Standardised scaling the values are bound between 0 and 1.
- Standardised scaling takes care of the outliers as they are capped at 1 whereas in Normalised scaling the outliers are scaled down but still exist.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Infinite value of VIF means that R-squared value = 1 indicating a perfect collinearity of that variable.
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   It is a way of finding out the distribution of a random variable whether it is exponential, uniform or normal. It is a scatter plot between quantiles.